SANDIA REPORT

SAND2013-0963
Unlimited Release
Printed January 2013

Nowcasting influenza outbreaks using open-source media reports

Jaideep Ray and John S. Brownstein

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation,
a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's
National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2013-0963
Unlimited Release
Printed January 2013

Nowcasting influenza outbreaks using open-source media reports

Jaideep Ray,
Sandia National Laboratories, P. O. Box 969, Livermore CA 94551

Prof. J. S. Brownstein,
Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115
{jairay@sandia.gov, John.Brownstein@childrens.harvard.edu}

Abstract

We construct and verify a statistical method to nowcast influenza activity from a time-series of the frequency of reports concerning influenza related topics. Such reports are published electronically by both public health organizations as well as newspapers/media sources, and thus can be harvested easily via web crawlers. Since media reports are timely, whereas reports from public health organization are delayed by at least two weeks, using timely, open-source data to compensate for the lag in "official" reports can be useful. We use morbidity data from networks of sentinel physicians (both the Center of Disease Control's ILINet and France's Sentinelles network) as the gold standard of influenza-like illness (ILI) activity. The time-series of media reports is obtained from HealthMap (http://healthmap.org). We find that the time-series of media reports shows some correlation (≈ 0.5) with ILI activity; further, this can be leveraged into an autoregressive moving average model with exogenous inputs (ARMAX model) to nowcast ILI activity. We find that the ARMAX models have more predictive skill compared to autoregressive (AR) models fitted to ILI data i.e., it is possible to exploit the information content in the open-source data. We also find that when the open-source data are non-informative, the ARMAX models reproduce the performance of AR models. The statistical models are tested on data from the 2009 swine-flu outbreak as well as the mild 2011-2012 influenza season in the U.S.A.

Acknowledgment

This work was supported by funds from the U. S. Department of Defense. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Contents

Figures

1 Introduction

Most national surveillance systems targeted at endemic diseases depend on passive or site surveillance of hospitalizations or monitoring of outpatient clinics. The data are generally collected via a network of physicians, and then compiled into reports by public health authorities. Due to many factors, including poorly funded collection efforts, these reports are not timely - in the best case (for example, the United States' Center for Disease Control, CDC), the reports are delayed by 2 weeks; in poorer countries, the delay is much longer [3]. Consequently, one turns to sources of data that may be outside of the clinical domain, but which can be recorded in a more timely manner. Examples include telephone triage calls [4], sale of over-the-counter drugs [5], school/work absenteeism [6] and online activity [7, 8, 9, 10]. Such data is meant to complement traditional surveillance methods, though results with respect to correlation and timeliness have been variable. Even if the signal in one data stream appears no earlier than in traditional surveillance techniques, the ability to collect a data stream (that is a proxy for traditional surveillance methods) in a near-real-time manner can allow early detection and response. The value of "predicting the present" for situations where data for the present may theoretically be available but not be accessible until the future is discussed in [11].

The use of online activity, as a proxy for disease activity, holds much promise. It operates under the assumption that prevalence of disease activity in a locale would lead to people endeavoring to find information on it, mostly via Web searches, in order to protect themselves or to seek treatment. Since symptoms of influenza-like illness (ILI) are well known, it was hypothesized that the frequency of influenza-themed searches would correlate well with disease activity [9]. This was further corroborated using the 2009 swine-flu outbreak in the USA [7] when the frequency of influenza-related online searches was used to predict ILI activity (as reported by the Center of Disease Control's network of sentinel physicians, ILINet). These predictions proved accurate. The method was also used in Sweden with encouraging results [8]. Further, since ILINet reports are delayed by two weeks, the predictions (strictly, nowcasts) proved very useful for tracking the waxing and waning of the pandemic in the United States. The same approach was tested for dengue in a number of tropical countries over 2003-2010 [12]. Timeliness of web query data is not an issue, and nowadays can simply be downloaded in real-time (from Google Flu [13] and Dengue [14] Trends). The same hypothesis has been applied to Twitter postings, with impressive results [15, 16, 17, 18, 19, 20].

However, surveillance methods that depend on social media and web searches are applicable only where Internet penetration is large, and where data from the search engine in question is easily accessible (e.g., Google is not the dominant search engine in China or Russia). Further, social media can be affected by sudden panics and rumors. It provides an unfiltered view into human behavior, which be influenced by factors other than disease activity. Reports of disease activity in the news media are less affected by such factors. They are published usually when they are newsworthy i.e., after a degree of fact-checking has been performed, and the outbreak is judged to have outstanding features. Since most media reports nowadays also appear online, this data can also be collected in a timely fashion. Similar reports of outbreaks, often lacking clinical details, are also submitted electronically (sometimes no more than an email) by volunteer networks of health workers to online fora. The potential of such reporting, called event-based biosurveillance (as opposed to tracking the indicators/symptoms of diseases, as is done in syndromic surveillance) is widely recognized and many efforts exist to collect such data [21]. They differ substantially in their details [22]. Most of these system include a web crawler (that fetches the documents) and a sophisticated text processing capability that identifies whether the document pertains to a discussion of a disease, and if so, its details (e.g., dateline of a newspaper article, the actual identity of the disease etc.). The text processing techniques used by these systems have been reviewed in [23]. By and large, these systems, sometimes called Digital Disease Detection (DDD), have been used for situational awareness [24].

One such system, HealthMap (HM, [2]), collects and indexes news articles, reports from public health authorities, as well as other sources of event-based outbreak-related data (e.g., ProMed-Mail [25]). While certainly not as comprehensive a reflection of disease-engendered human activity as Google searches or Twitter, it nevertheless serves as a "report aggregator", and provides a wide view of disease activity. This raises the potential of using HM data in much the same way as Google and Twitter data were used for nowcasting. Such a study has not been attempted to date.

HM data is quite different from Google or Twitter data. The latter is a direct reflection of human activity, some of which might be influenced by disease activity. In contrast, news media reports (which account for the bulk of HM data) are a result of a competition between multiple news reports, many of which may be far more important than disease activity. Since only a limited number of reports are published every day, disease-related reports may be sporadic, even though disease activity may prevail. Thus a time-series of the abundance of disease-related reports may be expected to be very noisy and bear a smaller correlation with a time-series of ILINet reports, as compared to Google searches and tweets. Other differences are more technical. Most newspapers simply reprint articles that are obtained from news agencies like Reuters i.e., articles from different newspapers may be identical and it is unclear whether they should be counted separately or together. Further, while a newspaper report might provide information about a specific outbreak e.g., swine-flu, a separate article may provide context i.e., describe its origins, without much reference to the outbreak. While no doubt germane to the swine-flu outbreak, it is moot if the publication of the article should be considered as a reflection of public interest in the outbreak. Details of how online reports are collected and categorized by HM are in [26].

In this work, we will explore if HM data can indeed be used for nowcasting. We will evaluate the correlation between time-series of disease-related media reports and ILINet data; we expect this correlation to be smaller than the 90% Pearson correlation coefficients observed for Twitter [17, 18]. We will explore time-series methods, based on autoregressive moving average models with exogenous inputs (ARMAX models), to nowcast disease activity i.e., ILINet data, with HM time-series acting as a guide ("exogenous input"). We will apply this at the country-scale (the US) where both the number of reports and the number of people detected with ILI symptoms may be expected to be large, and proceed down to city levels (New York City) to examine if the predictive skill of the ARMAX model shows substantial degradation. We will examine its performance during this 2009 swine-flu epidemic, when the disease dynamics were pronounced, and during the mild 2011-2012 influenza season, when both disease activity and the public interest in it were modest. Finally, we will consider a case where HM data is *misleading* i.e., has no correlation, to check the robustness of the model.

Below, in Sec. 2, we review the nowcasting models that use Google and Twitter data and develop the argument for ARMAX models. In Sec. 3, we develop the model, as well as data-smoothing specifications for HM data. In Sec. 4 we apply the model to different outbreaks and evaluate the performance of the model. In Sec. 5, we present our conclusions.

2 Literature review

A review of nowcasting models: Nowcasting of influenza epidemics gained widespread publicity when Google data was used to predict swine-flu dynamics 2 weeks ahead of CDC's ILINet data [7]. The authors modeled P, the percentage of ILI physician visits, as reported by CDC, as a function of Q, the ILI-related search query fraction, as $\text{logit}(P(t)) = \beta_0 + \beta_1 \text{logit}(Q(t)) + \varepsilon$, where ε is an error terms and

$$\text{logit}(X) = \log\left(\frac{P}{1-P}\right). \tag{1}$$

The form of the equation can be found in the original version of the paper [7]. Data was collated on a weekly basis. The coefficients β_0 and β_1 were computed by regression. The resulting model was found to have extremely high predictive skill. The challenge in model constructing lay in choosing the set of 45 keywords/phrases that identified whether a search bore any relevance to influenza. Data on influenza-related searches are now publicly available [13]. Recently, the set of keywords and the model itself was updated [27]. An older publication looked at the predictive capacity of Yahoo! searches containing "influenza" or "flu", but not "bird", "avian", "pandemic", "vaccine", "vaccination" or "shot" [9]. They found that the fraction of such searches $s(t)$ could predict the fraction of cultures testing positive for influenza ($c(t)$) with the model

$$c(t) = \beta_0 + \beta_1 s(t-x) + \beta_2 t + \varepsilon,$$

where x denotes the lag. The model with a 1-week lag performed best. When a model for mortality, rather than positive cultures, was created (with the same model form), a lag of 5-weeks was seen to be most predictive. In contrast a model of the form

$$O(t) = \beta_0 + \beta_1 Q(t) + \varepsilon,$$

was used in [12] to predict the *number* of dengue cases with Google query fraction Q. Thus, models have used both raw, logit-transformed and fractions as the dependent or predicted variable while the independent variable has always been a query fraction, with or without lags.

In [19], Eq. 1 was used to predict ILI activity using Twitter messages, rather than Google searches, as the independent variable. In contrast [17] used fractions without the logit transform, as did [16]. Thus, as in the case of Google searches as the independent variable, there is no consensus on whether a logit-transform is necessary; however, all models regress fractions (or rates) of ILI incidence on fraction of influenza-related Twitter messages.

Exploiting autoregression in the ILI datastream to improve the predictive skill of the resulting model has not been explored in any detail. While ARX models (autoregressive models with exogenous inputs) are quite common in econometrics [11], their use in nowcasting ILI activity using social media data has only been performed with Twitter as the exogenous datastream [28, 18]. Again, while the technical challenge primarily lay in detecting influenza-related tweets i.e., in text mining, an ARX model was considered. When applied to CDC ILINet data from the 2010-2011 season [18], it was found that a model that employed 2 weeks of lag in the dependent variable and none in the exogenous one (the fraction of influenza-related tweets, collated on a weekly basis) provided the most predictive skill. Interestingly, the dependent variable was a logit-transformed ILI fraction, whereas an earlier work by the same authors [28], applied to the tail-end of the swine-flu pandemic in 2009-2010 used ILI fractions i.e., without logit-transforms, and found that autoregression degraded the model's predictive skill. In both their works, the authors examined multiple models, and selected one using cross-validation. The dataset to which they regressed their model for 2009-2010 was small, and the autoregressive coefficients could have been removed due to over-fitting of the data.

A review of ARMAX models: Consider a time-series y_i, $i = 1 \ldots L$, which is deemed to be autoregressive (i.e., y_i it is correlated with its previous values $y_{i-j}, j > 0$) as well as correlated to an exogenous time-series x_i. In such a case, one may form a ARMAX model of the form

$$y_i = \alpha_0 + \sum_{j=1}^{N} \alpha_j y_{i-j} + \sum_{k=0}^{M} \beta_k x_{i-k} + \sum_{l=1}^{L} \gamma_l \varepsilon_{i-l}, \tag{2}$$

where ε is an error. One imputes multiple values of (M,N,L) and calculates $(\alpha_i, \beta_j, \gamma_l)$, the coefficients in the resultant model, by fitting to the learning-set data. This results in multiple "calibrated" models, the most predictive of which needs to be selected for further use. This may be detected via cross-validation (out-of-sample tests) as performed in [18] or, since the models are fully nested, one may compute various information criterion e.g., Akaike Information Criterion (AIC), Generalized Cross Validation (GCV), Portmanteau tests, RICE criterion, Final Predictive Error (FPE) and use them for model selection [29]. Theory on ARMAX models can be found in [30, 31]. Statistical packages, e.g. in R, implement ARMAX models and we will use Dynamic System Evaluation (`dse` [32]) for ARMAX modeling in this study.

3 Model formulation

Selection of modeling variables: In our model, we will use $x_i = \log_{10}(\hat{x}_i)$, as the independent variable, where $\hat{x}_i$ is the number of reports (news media, reports published by public health authorities etc) that have appeared during week i, as gathered by HM [2]. This is also referred to as the "exogenous input" (common in ARMAX terminology) as well as "HM data". This includes reports that deal with an influenza outbreak, as well as reports that provide context about the disease. The underlying hypothesis is that reports and reviews about a disease which are published during an outbreak reflect public interest in it, which in turn is proportional to the severity of the outbreak. Thus, for example, if a news article, originally from Reuters, is published verbatim multiple times in various newspapers, it is weighted by the number of times it appears, rather than once.

The exact definition of the dependent variable poses more of a challenge. One could consider $y_i^* = \log_{10}(\hat{y}_i)$ as the dependent variable, where $\hat{y}_i$ is the number (*not* percentage) of people detected with ILI symptoms during week i, as reported by a network of sentinel physicians. Such data can be downloaded from CDC [1], as well as its French equivalent, the French GPs Sentinelles network [33]. One could also consider y_i, the percentage of people visiting sentinel physicians who show ILI symptoms. This number, even during the swine-flu pandemic of 2009, did not cross 10% in the ILINet datastream. However, the number of patients seen by sentinel physicians can vary by 30%, and the number of data providers vary by a factor of two over a year, suggesting that y_i would be a better marker of disease activity than $\hat{y}_i$. However, it is also clear that due to the small values observed for y_i, its variation in time could be sensitive to the dynamics of diseases/symptoms other than ILI, and it is unclear which is a better measure of disease activity. We will first identify what measure of ILI activity should serve as the dependent variable to be modeled. We assume that all ILI symptoms indicate influenza, which is, of course, not strictly true.

In Fig. 1 (left) we plot y_i, y_i^* and x as functions of time. The data was obtained from CDC [1] and pertains to the US as a whole. We start with the first week of 2008 and present data for over 2 years i.e., the data contains the swine-flu pandemic of 2009 (the peak in Week 70–80) and the "normal" influenza season in the winter of 2009-2010 (the peak around Week 90–100). Note that the plot for y_i^* (blue line) shows that during the swine-flu weeks, the number of ILI patients did not actually cross the levels observed during the 2007-2008 winter influenza season. We see that the total number of patients (not just those with ILI symptoms) visiting the sentinel physicians (red crosses) remained roughly constant on a logarithmic scale, though the number of reporting physicians varied by a factor of 2 (red triangles). The percentage of ILI patients (red line) varied significantly and, from the CDC data, it is clear that it was smaller during the swine-flu outbreak than the preceding influenza season. This is largely a consequence of the expanded reporting during that period that garnered a lot of patients that did not have ILI symptoms. We see that the dynamic range of y_i is far larger than that of x_i, whereas that of y_i^* is comparable. In Fig. 1 (right) we plot the scatter plots of y_i and y_i^* versus x_i. It reveals that y_i^* is more correlated to x_i, compared to y_i; $C(x_i, y_i) = 0.534797, C(x_i, y_i^*) = 0.637895$. Further, the x_i time-series is a lot more jagged than the y_i and y_i^* series. Since it would be difficult to model a smooth function with a very rough one, it is clear that smoothing the x_i series will be necessary.

In keeping with previous literature, we will model y_i, and not y_i^*, since it captures the intensity of disease activity.

Matching spectral contents: We perform a Fourier decomposition of x_i as well as y_i via a Fast Fourier Transform, and normalize the magnitude of all Fourier coefficients by the largest one. These normalized magnitudes are plotted in Fig. 2. A smoothing kernel $\mathcal{K}(i, j)$ is applied repeatedly to the series x_i to obtain

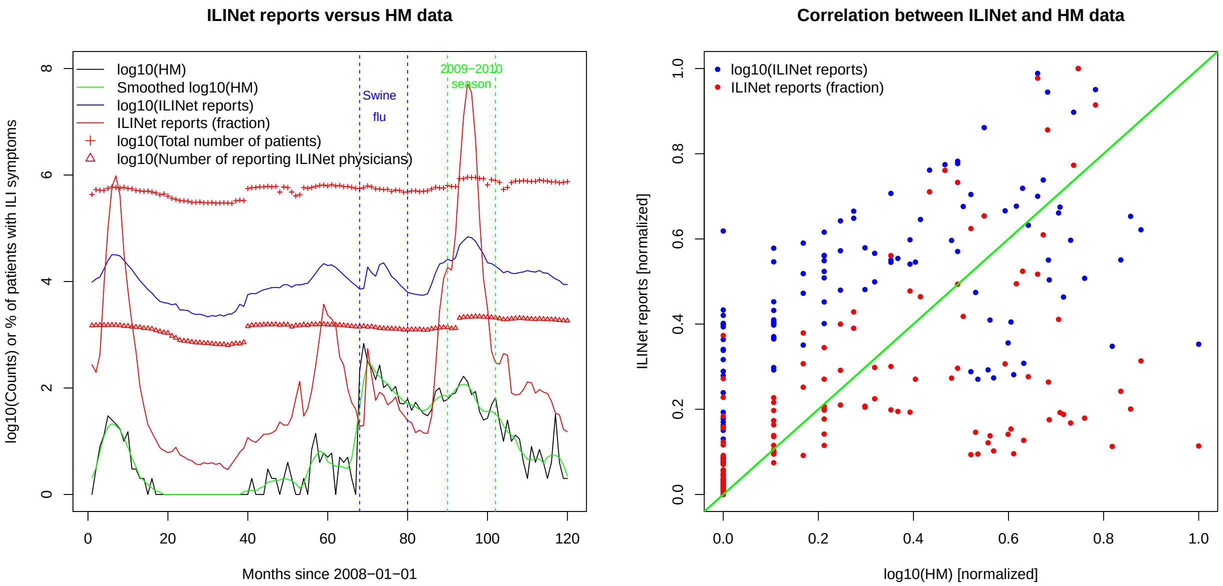

Figure 1: Left: Plots of x_i, y_i and y_i^* as a function of time, starting from the first week of 2008. "HM" stands for the weekly abundance of influenza-related reports gathered from HealthMap. Reports of patients with ILI symptoms are from CDC's ILINet. Red and blue lines correspond to y_i and y_i^* respectively. We see that all three time-series show some correlation. The green line plots the smoothed version of x_i, which has the same spectral content as (is as smooth as) the y_i time-series. The red crosses and triangles plot the evolution of the number of patients visiting sentinel physicians, and the number of reporting physicians over the period of interest. The blue dashed lines (Weeks 65–80) approximately demarcate the swine flu outbreak. The green lines demarcate the annual 2009–2010 flu season. Right: Scatter plot of y_i and y_i^* versus x_i. The green line shows perfect correlation. We see that for y_i, the red dots are scattered wider about the green line compared to the blue dots. The Pearson correlation coefficients are $C(x_i, y_i) = 0.534797, C(x_i, y_i^*) = 0.637895$.

its smoothed counterpart $x_i^{(s)}$,

$$x_i^{(s)} = \sum_{j=(i-2)}^{(i+2)} \mathcal{K}(i,j) x_j \quad \text{where } \mathcal{K}(i,j) = \frac{1}{C} \exp\left(-\frac{(i-j)^2}{2}\right) \tag{3}$$

and C is a constant so that the coefficients of the discretized smoothing kernel $\mathcal{K}(i,j)$ sum to 1. Thus, we smooth over a band of 5 weeks, distributed symmetrically about Week i. The normalized magnitudes of the Fourier coefficients, after one-, two- and three-applications of the smoothing kernel are also plotted in Fig. 2. We see that while the unsmoothed time series x_i has modes of high frequency with significant magnitude (and hence its rough nature), the high-frequency modes are quickly smoothed out. After three applications of the smoothing kernel, we see that the high-frequency content is essentially removed. We compare the the spectrum of y_i and the smoothed $x_i^{(s)}$ and find that smoothing once results in y_i and $x_i^{(s)}$ spectra that are most similar. We will henceforth drop the (s) superscript and use once-smoothed x_i in our work. The once-smoothed x_i is plotted as the green line in Fig. 1 (left).

Selecting the HM reports for inclusion in the x_i time-series: HM uses a dictionary-based process to classify documents and alerts fetched from the Web. The process automatically tags them as they enter the HM system. HM staff review the disease tags and make corrections when necessary. When a pattern of errors is noted, improvements are made to the dictionary that enable better disease classification. The

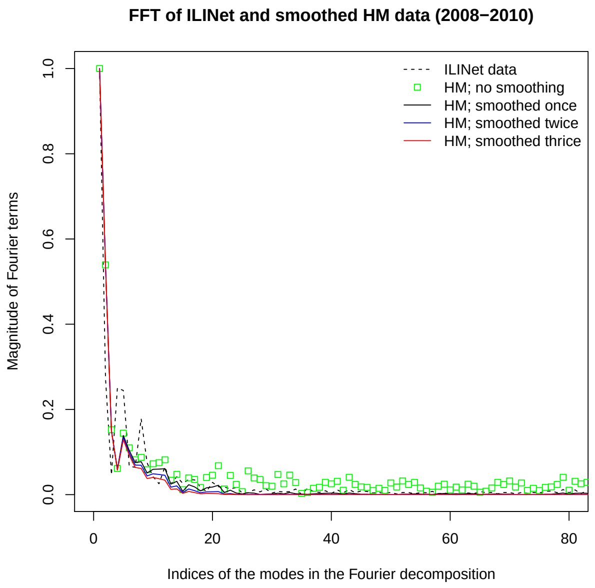

Figure 2: Magnitudes of the terms of the Fourier decomposition of y_i and x_i, after multiple rounds of smoothing. We see that the time series of y_i (dashed line) has negligible spectral power in the high frequency modes, whereas that of x_i (in green) has significant high-frequency content. As we smooth x_i to $x_i^{(s)}$ once- (black line), twice- (blue line) and thrice- (red line), the spectral content of y_i and $x_i^{(s)}$ become similar.

system tags all articles about seasonal influenza as influenza. During the first year of the H1N1 pandemic, the system often tagged news alerts as H1N1 since the articles were specifically about that subtype. The HM system currently tags over 250 different animal, human, and plant diseases.

The HM news feeds often pull in multiple versions of a similar story. For example, several cities may run articles based on an AP article. While they may include contextual information that is city specific, the heart of the story is identical. The system will automatically file these stories together. The first story that comes into the system is classified as the “primary” while other very similar stories are classified as “duplicates”.

The HM system receives a variety of articles and alerts on infectious disease-related topics. When the articles enter the system, they are parsed through a 5-way classifier that tags them as: Breaking, Context, Warning, Old News, or Not Disease Related. Articles with new information about current outbreaks are tagged as Breaking. Stories about policy, vaccine development, or basic research are tagged as Context. Warnings mean cases or an outbreak may be imminent but no outbreak exists currently. Old News covers stories about outbreaks that were old when the article was published; note that a Breaking story never becomes Old News because the tags are assigned relative to the publication date. Not Disease Related is used for such things as tornado outbreaks or “Bieber fever.”

In this work, we assume that the total weekly volume of discourse (via the media and on the Web) on a given topic would be indicative of the level of interest, which in turn would be dependent on the severity of the disease in question. Thus, the time-series that we analyze is composed of the sum of the weekly frequency of articles classified as Primary, Duplicates and Context, i.e., we analyze the total weekly volume of reports concerning an outbreak, or information about the disease.

Formulating the ARMAX model: We propose an ARMAX model, using Eq. 2 as its form. The fraction

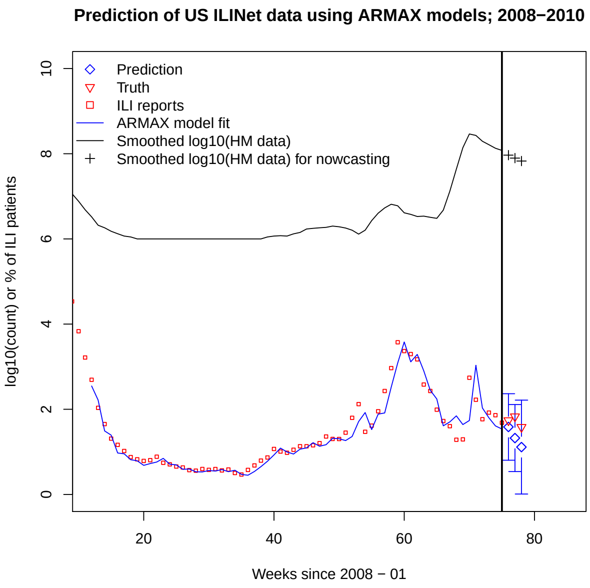

Figure 3: Demonstration of the forecasting capacity of the ARMAX model. The data used for calibrating the model are y_i (red symbols) and x_i (black line), obtained from [1] and [2]. x_i values for Week 76–78, in black crosses, are used to forecast ILI activity (blue diamonds with error bars denoting $\pm 3\sigma$ limits on predictive uncertainty). The true ILI activity, as reported by CDC's ILINet, are in red triangles (and overlap the blue diamonds).

of people with ILI symptoms, y_i, as reported by ILINet or French Sentinelles physicians, will serve as our dependent variable, to be forecast using once-smoothed time-series of HM data (news and other reports regarding the disease in question), x_i. The fitting of the model to data is performed using `dse` [32]. The maximum values of $M = N = L$ is set to 10 (unless otherwise mentioned) and an ensemble of models are calibrated to the (y_i, x_i) data. The models are first checked for stability. The final model is selected by the mechanism in `dse` which compares AIC, GCV, Portmanteau tests, RICE criterion, and FPE.

In Fig. 3, we illustrate forecasting using an ARMAX model. We plot y_i, from CDC's ILINet [1], using 78 weeks of data. The time-series data starts on the first week of 2008 and is plotted with red squares. The HM data (plotted as a black line) and y_i show a rise around Week 60, corresponding to the 2008-2009 winter influenza outbreak and another rise around Week 70, corresponding to the swine-flu outbreak. The rise in HM data around Week 70 corresponds to reporting on swine-flu. We have shifted the HM plot upwards by 6 for plotting clarity versus the y_i plot; x_i assumes values in (0, 2) rather than (6,8) as seen in the plot i.e., one has about a hundred influenza-related reports per week rather than in the millions. The learning set and testing set data are separated by a thick black line.

We construct an ARMAX model with 70 weeks of historical data starting from the point of forecast. Thus for forecasting Weeks 76–78, we use Weeks 6–75. 70 weeks are chosen so that the model, with a maximum lag of 10, has sufficient data for estimating the model coefficients and performing cross-validation tests. Note that when `dse` constructs the ARMAX model, it does so only with 70 weeks of data i.e, the 70 weeks of data are further partitioned to perform generalized cross-validation tests while choosing between competing models. The prediction of the training data by the ARMAX model is shown with a blue line. The model is then used to predict Weeks 76–78 (diamonds and error bars). The true CDC data is plotted with red triangles. The HM data that serves as an exogenous "signal" for predicting ILINet data beyond Week 75 are plotted with black crosses. It is clear that we have obtained a good prediction; further, the $\pm 3\sigma$ error bars are small (compared to the variability in y_i seen in the training data) and are thus informative. They are also seen to increase as the forecast horizon increases. While we allowed M, N, and L to vary up to 10, the final, selected model has $M = N = L = 4$. This also explains why the error bar for the third forecast week (Week 83) is far larger than the rest – the majority of the y_i values used by the ARMAX model to predict it are forecasts themselves.

To summarize, we can construct an ARMAX model for y_i, using x_i as the exogenous input. The size of the training set will depend on the availability of data, but in general we will attempt to have a training period of 70 weeks. The procedure for constructing the ARMAX model creates an ensemble of models and selects one based various metrics described above. The model finally selected is used to forecast y_i, the fraction of ILI patients as reported by sentinel physicians, three weeks ahead.

This page intentionally left blank

4 Tests

We will now explore the performance of the model, under various scenarios, to identify its limits of applicability. We will also compare its performance against purely auto-regressive (AR) models, (which does not incorporate HM data), to evaluate the impact of including an exogenous input (HM data) in our predictions.

We start with the 2009 swine-flu pandemic in the United States. In Fig. 4 (top left), the red line shows the % of patients visiting ILINet physicians (y_i), starting from the first week of 2008, collated on a weekly basis. We see the 2009 pandemic influenza outbreak around Week 70, and a higher peak during the winter of 2009-2010. The impact of the pandemic on online reporting and discussion is seen in the black plot, which is x_i, the logarithm (to base 10) of the number of reports on influenza or the pandemic as collected by HM weekly ("HM data"). Between Week 70–80 (April-June, 2009), x_i shows a steep rise. The plot has been shifted up by 6 i.e., there were about 100 articles/reports per week discussing influenza and allied topics during the period. We use the information contained in the ILINet and HM data (red & black lines) to perform forecasts at three points in time - Week 81–83, 97–99 and 111–113, as denoted by the thick black lines in the plot. Forecasts are performed for 3 weeks, using the ARMAX model (blue symbols) and an AR model (green symbols), fitted to y_i. The data from 70 weeks preceding the forecast dates were used to train and select the models, using the method described in Sec. 3. The ARMAX models used for forecasting Weeks 81–83 had a lag of 3 i.e. $M = N = L = 3$, whereas those for Weeks 97–99 and Weeks 111-113 had a lag of 2. The corresponding orders for the AR models were 4, 2 and 2. We also attempted to fit an ARMA (autoregressive moving average) and ARIMA (autoregressive integrated moving average) model to y_i; ARMA models were reduced to AR models during the fitting procedure and ARIMA were less accurate. It is clear that for the entire US, predictions using ARMAX have narrower $\pm 3\sigma$ error bars, as compared to AR models. Further, when the y_i time-series suddenly changes its slope, the mean predictions of the AR model (green triangles) show large errors (and the wrong trend), whereas those of ARMAX are closer to the truth in magnitude and trend. This is a consequence of being guided by the HM data x_i which is correlated with y_i. Thus, the net impact of assimilating x_i while forecasting y_i is to reduce predictive uncertainty.

We next consider a case where x_i is not a good guide to disease dynamics, and where the y_i signal is itself weak. The influenza season during the winter of 2011-2012 displayed such characteristics. We examine the performance of AR and ARMAX models for the Pacific HHS region. This is plotted in Fig. 4 (top right). The time-series spans 41 weeks, starting from the 40^{th} week of 2011. We see that y_i time-series does not show much structure, whereas x_i shows a decline i.e., there is very little correlation between the time-series x_i and y_i. Further, the % of patients with ILI did not reach 1% throughout the season. In this plot, x_i has been shifted up by 0.5 so that the y_i and x_i lines do not confound each other. We perform forecasts for Weeks 21–23 and 36–38, as indicated by the thick vertical lines. The lags of the selected ARMAX model were 1 and 3 for the two forecasts respectively; the corresponding numbers for the AR models were 1 and 7 respectively. The models were trained on 20 weeks of data preceding the forecasting date.We see that ARMAX (blue)and AR (green) models have similar behavior i.e., the ARMAX model identifies that the exogenous input x_i has little to contribute, and reduces to an AR model. Thus we see that the ARMAX model behaves in a predictable manner when the exogenous data proves uninformative.

The Pacific HHS region is rather large, and we next explore the performance of the ARMAX model in the smallest HHS region, New England. The plots are in Fig. 4 (bottom left). Here, the x_i time-series has been shifted upwards by 0.25 for plotting clarity. We see the same behavior as in the previous case - x_i does not provide any guidance towards disease dynamics, and the ARMAX model reduces to a AR model in performance. The lags for the ARMAX and AR models were 1 in both cases.

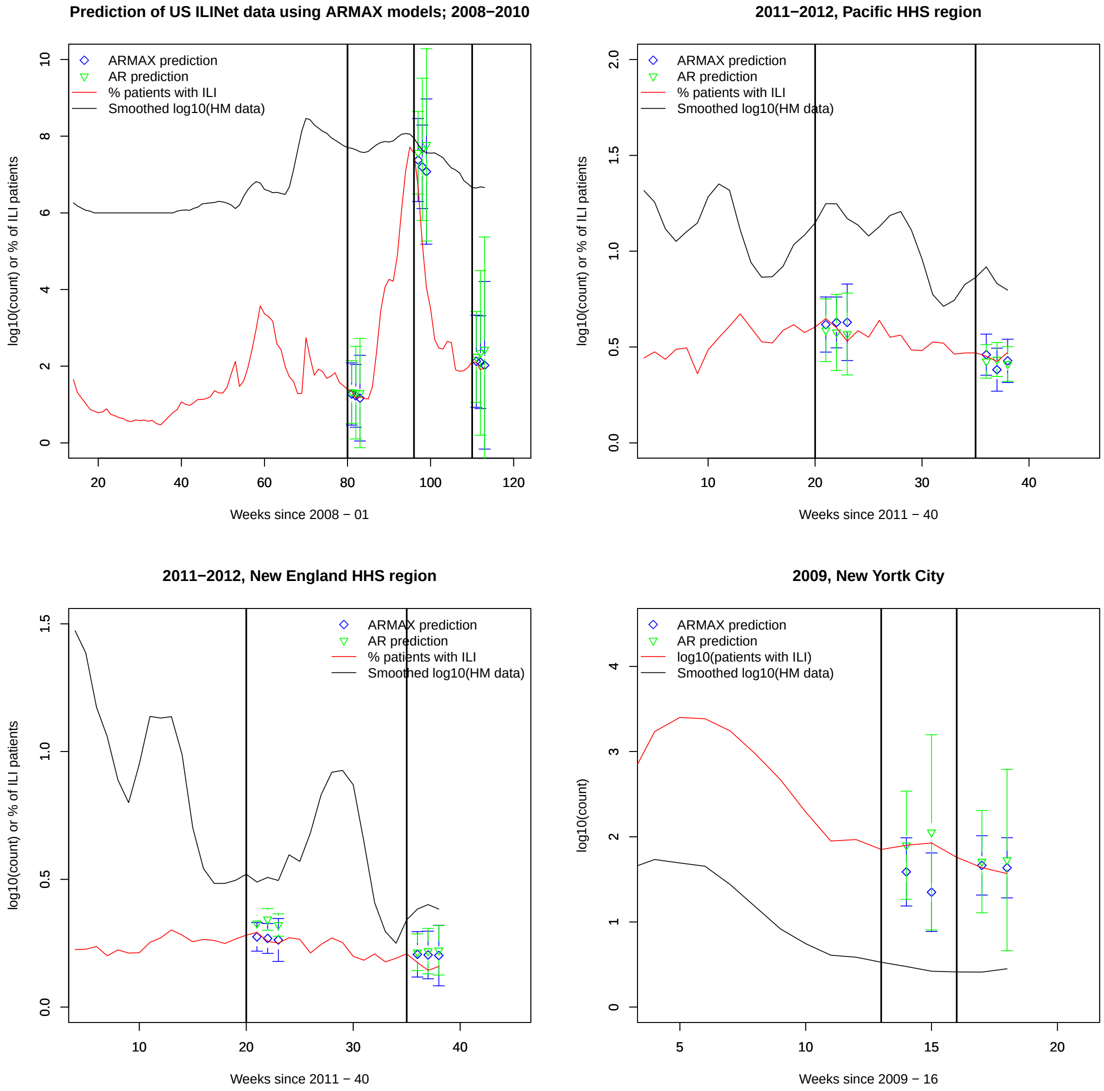

Figure 4: Performance comparison of ARMAX and AR models, with a view of evaluating the effect of including x_i in forecasting. The HM data consists of the log-transformed number of influenza-related reports (media and otherwise) and is plotted with a black line. The reports from sentinel physicians of CDC's ILINet are plotted in red. Plots show y_i, the % of patients detected with ILI by the physicians, except in the case of New York City, where we plot y_i^*, the log-transformed counts of ILI patients. Forecasts by ARMAX models are in blue; those by AR models in green. Top left: Results for US as a whole, during the 2009 swine-flu pandemic. ARMAX models have lower predictive error and the mean predictions are closer to the true data (in red) compared to AR models. Top right: We apply the same methodology to the 2011-2012 influenza season in the Pacific HHS region. The season was mild, there were few reports and there is little correlation $C(x_i, y_i)$. ARMAX models reduce to AR in performance. Bottom left: The same result as the top right figure is seen for New England, which has a smaller population and still fewer reports in the media. The weakness of y_i does not seem to have made any impact on the predictive skill of ARMAX. Bottom right: Predictions for New York City, during the swine-flu epidemic. Both the time-series x_i and y_i have lower numbers and have the potential to be noisy; however, they are strongly correlated. The correlation is detected and exploited by ARMAX models, giving it a far higher predictive skill than the AR model. The short training data does not seem to have adversely affected prediction accuracy, probably due to the strong $C(x_i, y_i^*)$ correlation.

Finally, we probe the behavior of the more complicated ARMAX model, versus AR, when the time-series itself is short (i.e., little training data). We consider the case of New York City, during the 2009 swine-flu outbreak. The data is plotted in Fig. 4 (bottom right). Here, we did not have access to y_i and modeling was performed using y_i^*. The data was obtained from [34]. The data set starts on the 16^{th} week of 2009. We performed two forecasts, for Weeks 14–15 and 17–18. All the available data before the forecast dates were used to train the ARMAX and AR models. We see that the behavior is similar to the first case (full US, during 2008–2010) where x_i was correlated to dependent variable. The net effect of including x_i is to reduce the predictive error, due to its correlation with y_i^*. Since the training set is so short, the maximum lag was limited to 3. The fitted models had lags of 1 and 2 (ARMAX) and 2 (AR, both cases).

Finally, we apply this modeling technique to data from France. We apply this method to the 2008-2010 influenza seasons, which included the swine-flu pandemic. During Week 65–75, we see a quick climb in media activity on influenza-related topics, after which it held steady for 25 weeks before declining. Disease activity, on the other hand, shows no such rise; it climbs steadily from its lows in summer to a maximum in winter (Week 100). Thereafter both media and disease activity declined in a correlated manner. Thus, it seems that the media activity was caused by swine-flu *elsewhere* and serves as a confounding information source. The challenge, therefore, is to verify if the ARMAX models identify x_i as not contributing any useful information, in which case they should perform like AR models. Disease activity data is obtained from [33]. In this case, we do not have information on y_i and the modeling is performed using y_i^*. The data is plotted in Fig. 5 (top). The red and black lines have their usual meaning. 70 weeks of training data were used for modeling, before producing the forecasts. The results show that the ARMAX and AR models have the same performance, though AR models have larger error bars. Thus, x_i is not contributing very much to the predictions. The ARMAX models have lags of 2, whereas the lags in the AR models vary between 3–6. In Fig. 5 (bottom), we zoom into particular predictions. In Fig. 5 (bottom left), during Week 41–43 and 81-83, the HM and Sentinelles data have very little correlation and ARMAX simply ignores x_i; hence the AR and ARMAX results are similar. In Fig. 5 (bottom right), during Week 119–121, the ARMAX model correctly identifies x_i to be informative and uses it. The AR model uses the shallower slope of the weeks before Week 119 and over-predicts the Sentinelles data, whereas as ARMAX follows y_i^*. During Week 131–133, x_i shows an opposing trend (y_i^* declines and x_i rises), leading ARMAX to slightly over-predict y_i^* whereas AR follows the y_i^* trend. However, the degradation due to x_i is minor. Thus, the ARMAX model behaved correctly in the face of confounding exogenous data; its performance reduced to that of an AR model when necessary. This robustness of ARMAX models has the potential to allow them to be used for forecasting with exogenous data, without explicitly checking for correlation between the predicted and exogenous variables.

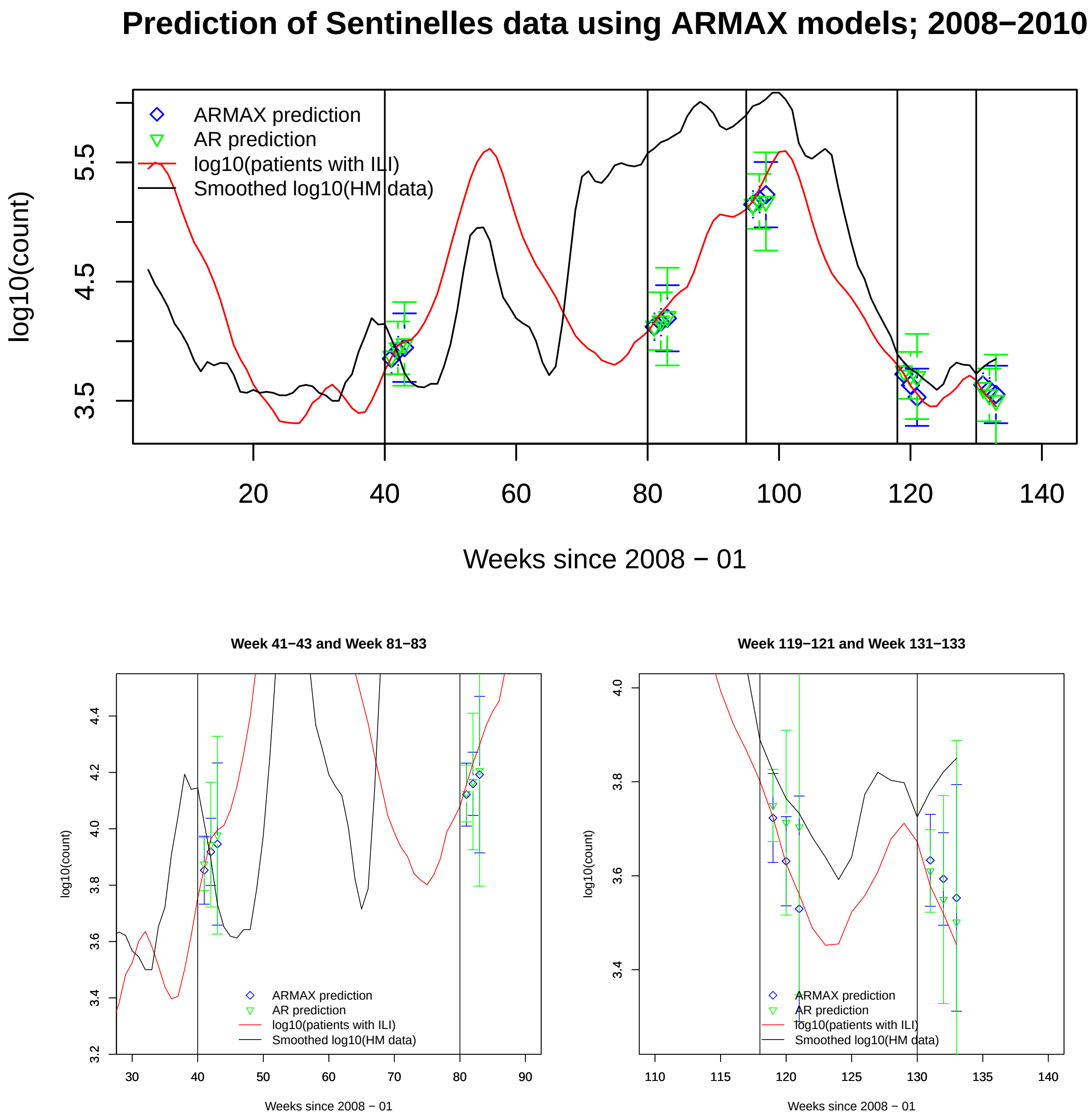

Figure 5: Performance comparison of ARMAX and AR models for influenza activity in France during the 2008-2010 influenza seasons. Above: We see that France largely escaped the pandemic (there is no sudden rise in disease activity [red line, plotting y_i^*)] around spring 2009, i.e. Week 70–80), though there was a lot of media activity regarding influenza (black line, plotting log-transformed time-series of the number of influenza-related news reports etc. from HM). ARMAX and AR predictions are plotted in blue and green respectively, and are similar i.e., the misleading media data is ignored by ARMAX models. Below (left): During Week 41–43 and 81-83, the HM and Sentinelles data have very little correlation and ARMAX simply ignores x_i; hence the AR and ARMAX results are similar. Below right: During Week 119–121, the ARMAX model correctly identifies x_i to be informative and uses it. The AR model uses the shallower slope of the weeks before Week 119 and over-predicts the Sentinelles data, whereas ARMAX follows y_i^*. During Week 131–133, x_i shows an opposing trend (y_i^* declines and x_i rises), leading ARMAX to slightly over-predict y_i^* whereas AR follows the y_i^* trend.

5 Conclusions

We have developed a method by which open-source reports of disease activity e.g., news media reports, can be used to nowcast influenza activity. This exogenous data was gathered from HealthMap (HM, [2]). We use a time-series of such reports, collated weekly, to predict reports of ILI activity gathered from sentinel physician networks. The quality of HM reports is far worse (shows smaller correlation with the time-series of ILI activity) than the data obtained from Web searches or Twitter. However, Web searches and social media data are obtained from specific companies e.g., Google and Twitter, and there are countries where their penetration is modest. Further, in poorer countries where the use of exogenous, open-source data would find more use (given poorly funded public health efforts) the Internet itself has a low penetration. Thus while conventional sources of social media data can perform disease nowcasting, they can do so in places where the need is not crucial. Due to good public health reporting, nowcasting/forecasting could, perhaps, be simply performed with auto-regressive moving average models. In contrast to Google and other social media data, news reports, as collected by HM, are quite plentiful even in poorer countries.

The poor quality of HM data does not make it less useful in nowcasting; it *is* information-rich. However, one requires more sophisticated methods to extract and exploit the information. The models used for nowcasting with Web searches and Twitter postings are simple linear models that exploit the tremendous correlation between social media and disease activity. In contrast, when using HM data, we rely on the autoregressive nature of disease activity and buttress it by assimilating exogenous, HM data. If this data stream is found to be informative, it is assimilated; if not, we fall back on autoregression. This is done naturally and automatically using ARMAX models. We have demonstrated them on disease activity data from the US and France. Some technical details, like the length of the training period for the models were not examined in detail, but we have identified certain lengths that work quite well. We have no reason to believe that the same training periods will not be sufficient when ARMAX models are applied to other countries.

The real nowcasting challenge lies in doing so for poorer countries where public health reporting displays longer lags behind disease activity. In such cases exogenous datastreams assume huge importance. It is likely that the reports that HM gathers will prove insufficient; rather one should use *all* exogenous data streams, like Web searches, social media activity, syndromic surveillance etc as independent predictors. They can then be assimilated using vectorized ARMAX models [35, 36]. State-space methods, to which ARMAX models can be reduced, have already been used to assimilate two exogenous data streams - humidity and Google Flu Trends - to predict and reconstruct influenza activity in New York City [37]. Thus ARMAX models (alternatively, state-space methods) allow a scalable approach to assimilating diverse open-source data, with a view of nowcasting in regions where conventional sources of public health reports are underdeveloped. In the process of doing so, the sensitivity matrices that the models calculate reveal much about the data-worth of competing open-source information streams.

This page intentionally left blank

References

[1] Fluview: National and regional level outpatient illness and viral surveillance. http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

[2] Healthmap webpage. http://www.healthmap.org/en/.

[3] E. H. Chan, T. F. Brewer, L. C. Madoff, M. P. Pollack, A. L. Sonricker, M. Keller, C. C. Freifeld, M. Blench, A. Mawudeku, and J. S. Brownstein. Global capacity for emerging infectious disease detection. *Proceedings of the National Academy of the United States*, 50:21701–21706, 2010.

[4] W. K. Yih, K. S. Teates, A. Abrams, K. Kleinman, and M. Kuldorff. Telephone triage service data for detecting influenza-like illness. *Public Library of Science, One*, 4, 2009. e5260.

[5] D. Das, K. Metzger, R. Hefferman, S. Balter, and D. Weiss. Monitoring over-the-counter medication sales for early detection of disease outbreaks – New York City. *Morbidity and Mortality Weekly Report*, 54:41–46, 2005.

[6] M. Besculides, R. Hefferman, F. Motashari, and D. Weiss. Evaluation of school absenteeism data for early outbreak detection. *BMC Public Health*, 5:105–105, 2005.

[7] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, and M. S. Smolinski. Detecting influenza epidemics using search query data. *Nature*, 457:1012–1014, 2009. Original version at http://research.google.com/archive/papers/detecting-influenza-epidemics.pdf.

[8] A. Hulth, G. Rydevik, and A. Linde. Web queries as a source for syndromic surveillance. *Public Library of Science, One*, 4, 2009. e4378.

[9] P. M. Polgreen, Y. Chen, D. M. Pennock, and F. D. Nelson. Using Internet searches for influenza surveillance. *Clinical Infectious Disease*, 47:1443–1448, 2008.

[10] H. A. Johnson, M. M. Wagner, W. R. Hogan, W. Chapman, and R. T. Olszewski. Anlysis of Web access logs for surveillance of influenza. *Studies in Health Technology and Informatics*, 107:1202–1206, 2004.

[11] H. Choi and H. Varian. Predicting the present with Google Trends. *Economic Record*, 88:2–9, 2012.

[12] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein. Using Web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. 5(5), 2011. e1206.

[13] Google Flu Trends. http://www.google.org/flutrends/.

[14] Google Dengue Trends. http://www.google.org/denguetrends/.

[15] M. J. Paul and M. Dredze. You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2001.

[16] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of Twitter to track disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *Public Library of Science, One*, 6(5), 2011. e19467.

[17] V. Lampos and N. Cristianini. Nowcasting events from the Social Web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4), 2012. Article 72.

[18] Harshvardhan Achrekar. *Online social network flu tracker: A novel sensory approach to predict flu trends*. PhD thesis, University of Massachusetts, Lowell, 2012.

[19] Aron Culotta. Detecting influenza outbreaks by analyzing Twitter messages. arXiv:1007.4748 [cs.IR].

[20] Aron Culotta. Ligthweight methods to estimate influenza rates and alchohol sales volume from Twitter messages. *Language Resources and Evaluation*, 2012. doi:10.1007/s10579-012-9185-0.

[21] D. M. Hartley, N. P. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. P. Linge, A. Mawudeku, N. Collier, J. S. Brownstein, G. Thinus, and N. Lightfoot. Landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3, 2010. doi:10.3134/ehtj.10.003.

[22] M. Keller, M. Blench, H. Tolentino, C. C. Freifeld, H. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein. Use of unstructured event-based reports for global infectious disease surveillance. *Emerging Infectious Diseases*, 15(5):698–695, 2009.

[23] Nigel Collier. Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health: An International Journal for Research, Policy and Practice*, 7(7):731–749, 2012.

[24] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital disease detection – Harnessing the Web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009.

[25] ProMed-Mail webpage. http://www.promedmail.org/.

[26] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. HealthMap: Global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association*, 15:150–157, 2008.

[27] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *Public Library of Science, One*, 6(8), 2011. e23610.

[28] H. Achrekar, A. Gandhe, R. Lazarus, S-H Yu, and B. Liu. Predicting flu trends using Twitter data. In *Proceedings of the First International Workshop on Cyber-Physical Networking Systems*, 2011.

[29] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, pages 40–79, 2010.

[30] G. C. Goodwin and K. S. Sin. *Adaptive Filtering, Prediction and Control*. Prentice-Hall, Inc, 1984.

[31] Lennart Ljung. *System Identification: Theory for the user*. Prentice-Hall, Inc, Upper Saddle River, NJ, 1999.

[32] Paul Gilbert. Dynamic System Estimation (time series package). http://cran.r-project.org/web/packages/dse/index.html.

[33] Sentinel network, INSERM, UPMC. http://www.sentiweb.fr.

[34] Epi Data Brief. http://www.nyc.gov/html/doh/downloads/pdf/epi/databrief5.pdf, 2009. Indexed on the page: http://www.nyc.gov/html/doh/html/episrv/epidata.shtml.

[35] Helmut Luetkepohl. Econometric analysis with vector autoregressive models. Economics Working Papers ECO2007/11, European University Institute, 2007.

[36] Eric Zivot and Jiahui Wang. *Modeling Financial Time Series with S-Plus*. Springer, New York, 2003.

[37] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Science of the United States*, 109(50), 2012. doi:10.1073/pnas.1208772109.

DISTRIBUTION:

1	Jaideep Ray, 08954	MS 9159
1	Technical Library, 08944 (electronic)	MS 0899

Sandia National Laboratories